NPS-OR-97-019

NAVAL POSTGRADUATE SCHOOL
Monterey, California

Dynamic, Interactive Statistical Research Papers on the Web

by

Samuel E. Buttrey
Gordon H. Bradley

December 1997

Prepared for: Air Force Office of Scientific Research
Bolling AFB, DC 20332-0001
and
Office of Naval Research
Arlington, VA 22217

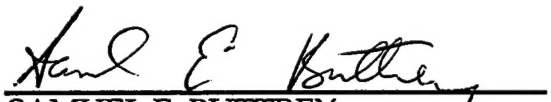
NAVAL POSTGRADUATE SCHOOL
MONTEREY, CA 93943-5000

Rear Admiral M. J. Evans
Superintendent

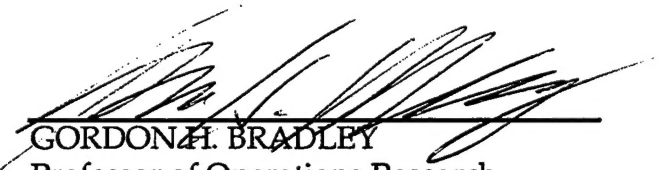
Richard Elster
Provost

This report was prepared for and funded by the Air Force Office of Scientific Research, Bolling AFB, DC 20332-0001 and the Office of Naval Research, Arlington, VA 22217.

This report was prepared by:

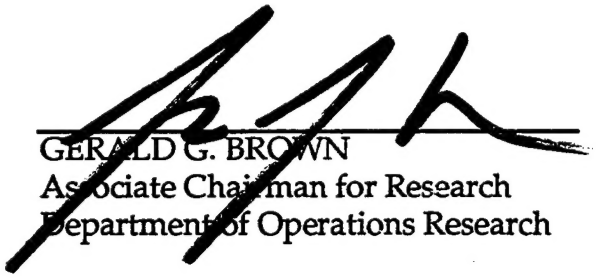


SAMUEL E. BUTTREY
Assistant Professor
of Operations Research

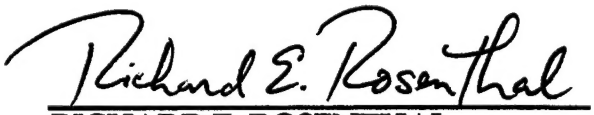


GORDON H. BRADLEY
Professor of Operations Research

Reviewed by:



GERALD G. BROWN
Associate Chairman for Research
Department of Operations Research



RICHARD E. ROSENTHAL
Chairman
Department of Operations Research

Released by:

DAVID W. NETZER
Associate Provost and Dean of Research

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	22 December 1997	Technical

4. TITLE AND SUBTITLE
Dynamic, Interactive Statistical Research Papers on the Web

5. FUNDING NUMBERS
N0001498WR20001

6. AUTHOR(S)
Samuel E. Buttrey and Gordon H. Bradley

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Postgraduate School
Monterey, CA 93943

8. PERFORMING ORGANIZATION REPORT NUMBER
NPS-OR-97-019

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Office of Scientific Research, 110 Duncan Avenue, Suite 100, Bolling AFB, DC 20332-0001
Office of Naval Research, 800 North Quincy Street Arlington, VA 22217

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Statisticians use the Internet and the World Wide Web to exchange e-mail and to view, retrieve and distribute research papers, data sets, source code and other materials. In addition to the text, data, and graphs in paper reports, electronic papers can include sound, animation, video and other computer-based media products that can improve the presentation of the results. Recently it has become possible to embed executable content --- programs --- into electronic documents. A statistician can now write computer programs to perform calculations described in the paper, embed the code in the paper, and distribute the paper over the Internet. Any reader can execute the code locally and interact with the program. This capability transforms a static presentation of research results into a presentation that can be dynamic and interactive.

This paper describes how the new capability can change the way statistical research is conducted. An illustration is given from the vision research field. In this real-life problem, researchers need to estimate the parameters of the ellipses that best fit particular data sets. Our software allows any researcher with a browser to compute the parameters, get estimates of standard errors, and draw pictures to illustrate the quality of the fit.

14. SUBJECT TERMS	15. NUMBER OF PAGES
World Wide Web; Embedded Algorithms; JAVA; Ellipses	10
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

Dynamic, Interactive Statistical Research Papers on the Web

Samuel E.Buttrey, Gordon H. Bradley, Naval Postgraduate School
Samuel E. Buttrey, Code OR/sb, Naval Postgraduate School, Monterey CA 93943

Key Words: World Wide Web, Embedded Algorithms, Java, Ellipses

1. Introduction

Statisticians use the Internet and the World Wide Web (WWW) to exchange e-mail and to view, retrieve and distribute research papers, data sets, computer source code and other text materials. In addition to the text, data sets, and graphs that are in paper reports, electronic papers can include sound, animation, video and other computer-based media products that can improve the presentation of the results. Recently it has become possible to embed executable content — programs — into electronic documents. A statistician can now write computer programs to perform calculations described in the paper, embed the code in the paper, and distribute the paper over the Internet. Any reader of the paper can execute the code on his or her computer and can interact with the executing program. This capability transforms what was once a static presentation of research results into a presentation that can be *dynamic* and *interactive*. This paper describes how the new capability can change the way statistical research is conducted. An illustration is given from the field of vision research. In this real-life problem, researchers need to find the best-fitting ellipses for particular sets of data that they gather. Our software allows any interested researcher with a browser to compute the relevant parameters, get estimates of their standard errors, and draw pictures to illustrate the quality of the fit.

2. Dynamic, Interactive Documents

The most widely used technology for embedding executable code into electronic research papers is the Java programming language. A compiled Java program can be attached to a Web page and then downloaded over the Internet and executed on the user's machine using a Java-capable browser. Most browsers, including the most widely used ones, are Java-capable. Java is a full-featured object-oriented programming language that was designed to execute without modification on a variety of different computers, including PC's, Apple computers, and most Unix machines. Although Java is not the only technology to be embedded in Web documents and distributed over the Internet, it is the most widely used and widely supported.

Under the control of the reader, a Java program can be executed on the reader's computer. This "dynamic" execution can produce the full range of computation and presentation of results possible on modern windows-based computers. This includes animation, sound and full-color display of results. In addition, the reader can "interact" with the running program to control its execution and to supply data.

2.1 Embedded Algorithms

The "dynamic" and "interactive" capabilities are both important to the propagation of statistical research, but it is the second that is perhaps the more vital. It is common practice today for a researcher to produce an algorithm or technique, to try it out on some data, and to publish the algorithm as part of a research paper. In some cases the reader is shown neither the full code that implements the algorithm, nor the data on which it was run. In others, the code is available only in a form that many readers cannot use. Even in the most benign case, it generally falls to the reader to "port" the code to his or her environment, and to validate the algorithm with his or her own data.

With the recent advances in Internet technology, an algorithm can be embedded inside a research paper. This allows the reader not only to see the results of the algorithm as it applies to the author's data, but, given access to the data, to reproduce those results. This allows immediate reproducibility of the author's computational results.

Even more importantly, an embedded algorithm can permit the reader to use it on his or her own data. This, after all, is often the reader's motivation: to use the newly-described technique to solve one of his or her problems. By applying the algorithm to

Prof. Buttrey is supported by the Research Initiation Program of the Naval Postgraduate School. Prof. Bradley is supported by the Air Force Office of Scientific Research and by the Mathematical Sciences Division, Office of Naval Research.

his or her own data, the reader can evaluate the robustness of the procedure as well. All of this is made instantly accessible to any user who has the use of a Web browser. Readers who are from other disciplines — who are less often able to "port" code and who are frequently in position to benefit most from the fruits of the research — can find these prewritten algorithms particularly useful.

It is certainly applied papers that can benefit most from technologies that permit dynamic visualization and from the ability to embed algorithms. There will always be a place in statistics for research of a theoretical nature that requires no pictures or computer programs to show its results. Still, these papers, too, can benefit from the new technologies. In addition to "links" to whatever data might be used in examples, a paper might maintain links to reference materials, other papers, and statistical encyclopedias, for example, for definitions and examples of terms that might not be in the casual reader's vocabulary. Every researcher has faced, in reading a research paper, the task of locating reference papers that cannot be found in the local library; we envision a world in which most of those papers are one mouse-click away. The emerging Internet technologies can, and should, change the way that statisticians do research.

2.2 Java and Its Rivals

Of course none of these advances require the use of the Java language specifically. However, this language has proven itself to be well-adapted to Internet concerns through its built-in network awareness and inherently secure design. Some drawbacks remain in these areas. Currently (as of October 1997), Web browsers do not allow, for security reasons, embedded Java programs to read or write to files on the reader's machine. This is not inherent to the Java language nor to Java's security model; the next generation of browsers will allow a reader to control access to the files on his or her machine, allowing a Java program to access data from, and write results to, particular local files or files in particular directories. Furthermore, the language has been slow. Initially, Java programs were one-tenth to one-twentieth as fast as C, C++, or FORTRAN programs performing the same calculations. Recently, new compiler technology has reduced this disparity and has the potential for making Java programs as fast or faster than other languages. Development time in Java is generally reduced because of the language's simple nature as embodied, for example, in the absence of pointer arithmetic.

2.3 The Future of the Research Paper

The primary medium to report and to preserve research results has been the research paper. We expect that this will continue. In recent years there has been a great increase in "electronic publication"; it was estimated in May 1995 that there were some 300 electronic research journals (Economist (1995)). The obvious advantage of an electronic publication is the speed with which it can be widely disseminated. However, part of the reduction in distribution time has come from avoiding the refereeing process that is so critical to the advancement of knowledge. This avoidance is certainly not inherent in the technology, and refereeing procedures can and have been set up to certify the quality of publications. After the publication has been refereed, electronic distribution is instant. Journal publication requires additional effort to prepare for publication (possibly the creation of galley proofs and additional proofreading) and then a wait in a publication queue that may be months or years long. However, once an article has appeared in a journal it has permanence — it will always be available — and it has stability — it will be the same to all readers.

In order for electronic publication to be acceptable for archival publication of research there must be procedures to protect the existence and integrity of the publication. For existence the author and the professional community must be assured that the publication will be easily and cheaply available to scholars for at least hundreds of years. A widespread computer virus could produce the electronic equivalent of the burning of the library of Alexandria. There must be additional procedures that assure that the publication continues to exist in its original form; there can be no undocumented revisions.

Assuming that the primary requirements of existence and integrity are met, we can consider additional benefits. The publication could be made available at very low cost by a provider who would supply only electronic access; the user would bear any cost of transferring it to paper. In 1991 researchers at Los Alamos National Laboratory set up a service to store and distribute electronic research preprints. It has been reported (Economist (1995)) that they handle 40,000 requests per day from 20,000 researchers. Electronic publications can provide instant access to references that are also available on the Internet. While the economics of electronic publication have yet to be worked out, it is to be hoped that this form of distribution will be substantially less expensive than paper journals, and therefore that these publi-

cations can be made more widely available than the vast and expensive set of journals being published today.

Electronic authors could shorten or eliminate introductory materials by linking to, rather than including, them. This can be carried one step further by imagining that communities of researchers will build Internet resources for each research field that can provide common materials: for example, a simple introduction to the problems, a set of standard notation and definitions, a set of standard test problems, a comprehensive bibliography, etc. As these resources change over time they would have to be dated so that a researcher linking to a publication written five years ago would get the list of references that existed at the time the publication was written. Electronic publications can also easily attach corrections and additions to the publication (while preserving the integrity of the original publication). This encourages other researchers to maintain links to the publications rather than saving a copy that becomes static the moment it is copied.

The present technology typically has a one-way link from a publication to a publication that it refers to. A two-way link could be maintained that keeps with each publication links to other publications that are linked to it. This might be accomplished by a process where a new publication "registers" with some or all of the existing publications it links to. This makes explicit the dynamic web of publications that makes up the body of research results in a specific area.

Links to other documents are valuable in research papers; however, hypertext does not have as much impact on research publications as on some other kinds of publications. In contrast to publications that present an interrelated body of knowledge that can be traversed in any number of different ways, research publications are essentially a linear presentation of material to convince the user about some proposition. They are more like a legal brief than a walk through a museum or an encyclopedia. In the jargon of hypertext, research publications are inherently linear and links are primarily used to connect to materials outside the publication rather than allowing the user to read the publication in a nonlinear fashion.

3. Example

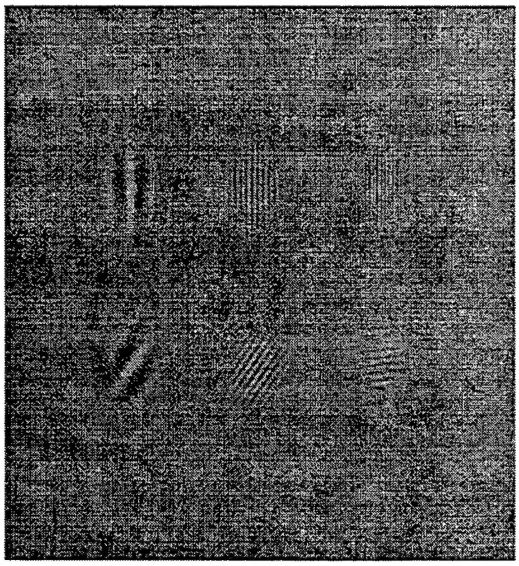
3.1 The Chromatic Oblique Effect

In this section we give an example of a real-life problem whose solution uses these technologies. In the vision research community, it is well known that the human eye exhibits an "oblique effect." That is, (monochromatic) stimuli are perceived differently depending on whether they are oriented horizontally or vertically on the one hand, or at 45 degrees or 135 degrees to the horizontal on the other (see, for example, Sellers et al. (1986)).

What is not known is whether this oblique effect persists in color vision. Experiments thus far seem to have been inconclusive. These experiments are difficult to design and carry out, because the colored stimuli need to be presented at equal luminance. (Otherwise, of course, the change in sensitivity might be equally well ascribed to changes in colors or to changes in luminance.) Since luminance is a fact of perception, different subjects have different ideas of what constitutes equiluminant stimuli.

A colleague of ours is gathering data to try to determine whether the chromatic oblique effect exists. In his experiment a subject looks at a screen and stimuli are projected onto it. Each stimulus consists of sets of parallel lines, oriented either at 0 degrees or 90 degrees ("non-oblique") or at 45 or 135 degrees ("oblique"). The parallel lines (which are presented at different thicknesses in different experiments) are alternately red and green. Ideally these lines would be presented as equally luminant; that way, the subject would need to use color vision to differentiate the parallel lines and to detect the stimulus. This is difficult to do, despite a certain amount of subject-specific calibration that takes place before the experiment begins. So in this experiment the stimuli are known to be nearly but not exactly equiluminant.

Figure 1: Grid Stimulus



The background on which the parallel lines are projected consists of a constant color that is produced by having each of the red, green, and blue electron guns in the monitor operate at half power. The stimuli are produced across the range of power available to the red and green guns.

These power readings are standardized so that half-power is set to 0, no power is -1, and full power is $+1$. Then each stimulus can be plotted on a graph where the horizontal axis is the red stimulus and the vertical is the green. Any ray out from the origin then depicts a set of constant red/green contrasts, where the luminance increases with distance from the origin but there is no change in chromatic contrast. (In the "negative" directions, the removal of red or green power increases the ability of the subject to see the stimulus in contrast to the background.) Figure 1 shows (in monochrome) examples of stimuli with different orientations (top and bottom) and different thicknesses or spatial frequencies (left to right). Incidentally, the ability easily to provide high-quality color graphics is another advantage of a network-based rather than paper-based journal article.

The subject is asked to indicate whether the stimulus is visible or not. At the origin, the stimulus is never visible because its luminance is zero and it does not contrast with the background. Far from the origin, the luminance is large, and the stimulus is always visible. The object of the experiment is to determine the thresholds: the critical values of the luminance, at which the stimulus is visible but below which it cannot be seen. Vision theory suggests that the set of these thresholds, plotted in this red-green space, should form an ellipse centered at the origin. The statistical problem here, first, is to fit an ellipse to this set of thresholds, and to be able to construct confidence intervals and test hypotheses about the parameters of the ellipse, and, second, to make this information available to non-statistical users.

3.2 Fitting Ellipses

Fitting ellipses to noisy data is a problem that has been considered for some time. This problem arises most notably in the pattern recognition world, where the object is to recognize a circle that is presented to the viewer at an angle. Among the early works on this problem is Bookstein (1979); the recent article by Rosin (1993) on data normalization includes a number of useful references.

Previous work on this problem has either not assumed a specific error structure or has chosen one incompatible with the current example. In the current problem, the errors in the observations are clearly radial, since each set of stimuli make up a ray out from the origin of constant chromatic contrast. (In contrast, common error models assume either that the x and y co-ordinates of the data are subject to independent errors, or that the error is perpendicular to the ellipse at that point, which is not the same as being radial.)

A second consideration is that previous investigators have mostly considered the general conic section equation

$$Axy + Bx^2 + Cy^2 + Dx + Ey + F = 0$$

in their model of the ellipse. Our interest lies more in the familiar parameterization where an ellipse parallel to one of the co-ordinate axes satisfies

$$(x - x_0)/a^2 + (y - y_0)/b^2 = r^2$$

because it is the semi-axes a and b about which we would like draw inferences.

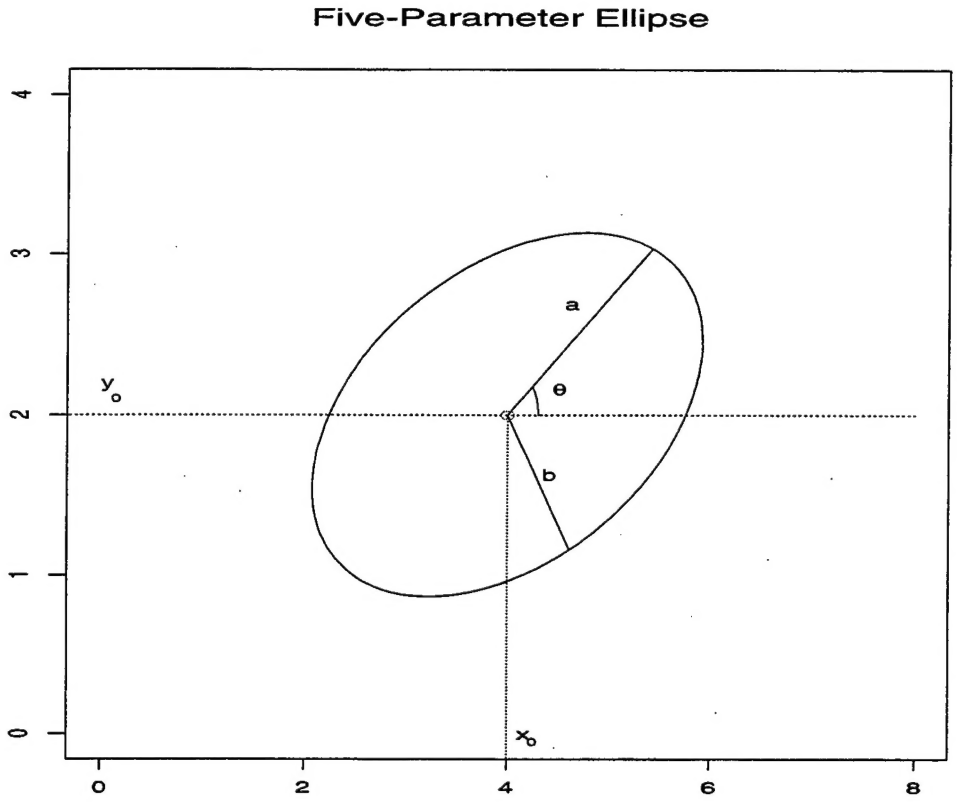
Let us write our data, $\{x_i, y_i\}, i = 1, \ldots, n$ in polar co-ordinates as $\{r_i, t_i\}, i = 1, \ldots, n$. Let x_0, y_0 be the ellipse's center (this is often assumed to be the origin, but that assertion is something to be tested), and use θ to designate the angle the ellipse makes with the positive x-axis. (Figure 2 shows an ellipse with its parameters in our parametrization indicated: a and b are the semi-axes, θ is the angle of inclination, and (x_0, y_0) are the co-ordinates of the center.) As usual, a "hat" designates a fitted value. Then our model can be simply stated as

$$r_i = \widehat{r}_i + \epsilon_i, \epsilon_i \text{ i.i.d. } \mathrm{N}(0, \sigma^2).$$

Suppressed in this notation is the dependence of the fitted $\widehat{r}_i$ on the underlying parameters a, b, θ, x_0, y_0. It can be argued that the errors can be expected to be larger where the ellipse is wider; perhaps a more appropriate model would assume a constant coefficient of variation for the ϵ. For the current purpose the assumptions of Normality and equal variance will simply be left as assumptions.

As in ordinary linear regression, the maximum-likelihood solution under this model coincides with the least-squares solution. However, the model is non-linear in the parameters, so some iterative technique is required to find the best-fitting set of parameters. For this purpose we constructed a non-linear regression routine in Java so that these ellipses can be fit by both our local colleague and by other researchers in other facilities with different machines and software. (Unhappily, our system uses the latest version of Java, version 1.1.2, which is not yet supported by most browsers. However an application

Figure 2: Geometry of Ellipse



version — a full executable program, not an applet that runs inside a browser — is available from the authors by e-mail for those users who have installed the Java Run-Time Environment.)

4. Results

It is not possible to show in this journal article the full results of our labor: this is part of our point, that this labor has produced an algorithm that can be used instantly and need not be re-created by interested users. However a couple of results are interesting. After having written the code to find the estimates of the ellipse parameters, it remains to find their standard errors. These, of course, will be required to make any sort of statement about whether two ellipses (an oblique and a non-oblique one, for example) are "significantly different."

One usual approach to this problem would be to rely on some asymptotic result. It is well known that asymptotically (under some regularity conditions) the vector of maximum-likelihood estimates is multivariate Normal with expected value given by the true values of the parameters. Furthermore, the covariance matrix of the estimates is given by the inverse of the Fisher information matrix (Lehmann (1983)). The Hessian — the matrix of second derivatives of the log-likelihood — is often used to approximate the Fisher information, and this matrix is easy to obtain. Our experience shows, however, that this small-sample approximation is not an accurate one for the sample sizes in our problem (generally fewer than 80 observations).

A second solution, and perhaps the most attractive, would come from simulation, either starting with the estimates and generating a set of errors from the postulated error distribution or, in larger samples, estimating the standard errors by the bootstrap. The simulation solution is costly from a computational standpoint, but under the Internet model this cost is borne by the users of the software, not by the developers, since it runs on the local machine.

A third choice, and one that our research has so far shown to be reasonable, is to rely on the well-known result that -2 times the log of a likelihood ratio has, asymptotically, the chi-squared distribution.

The software uses this result to examine the hypothesis that additional parameters are not needed in the model, and the associated confidence intervals consist of all points in parameter space for which this test does not reject that hypothesis. So for example we can examine the null hypothesis that the true center of a particular ellipse is in fact at the origin by fitting a five-parameter ellipse and comparing that likelihood to the likelihood under the three-parameter model in which the center is fixed at $(0, 0)$. Under the null hypothesis, -2 times that likelihood ratio has the χ^2_2 distribution. Our software will compute this statistic and report it and its associated p-value. In fact it will also draw a confidence contour on the plot of the thresholds so that the investigator can see clearly the extent to which the ellipse is off center. (This calculation has led to the conclusion that one monitor used in the experiment has not been properly calibrated.)

The hypothesis that two ellipses (one based on oblique stimuli, the other on non-oblique stimuli) are not significantly different can be tested by comparing the model that consists of two separate ellipses (and which therefore totals ten parameters) to the restricted model in which all the data is presumed to come from a single ellipse (a model with five parameters). A nested F-test similar to the usual linear regression F-test produces a statistic that is a good approximation to the F (Seber and C. Nelder (1989)). As with the χ^2 approximation, our simulation results have so far borne out this F approximation as believable.

Another way to compare data from oblique and non-oblique stimuli is to specify that the two ellipses differ only in a single parameter: that oblique stimuli and non-oblique stimuli produce ellipses with the same values of b, θ, x_0 and y_0 but that the two values of a differ (this is the oblique effect in action). Again an F-test can test this hypothesis.

5. Conclusions

Statisticians can now construct and distribute over the Internet dynamic, interactive electronic research papers that include embedded computer programs that can execute immediately on the reader's computer. This can allow the reader to reproduce the author's results and to immediately apply the embedded algorithms to his or her own data. The Java program for fitting ellipses to vision data illustrates how a sophisticated statistical analysis can be made available instantly to vision researchers around the world. Readers need not know how to port or compile code; the research paper not only describes the steps needed to perform the computations but actually performs them as well. This ability to include algorithms in papers brings an immediacy to the author-reader interaction that has not been possibly previously. There is no limitation on the type of computer or software that a reader needs (other than that it be equipped with one of the popular Web browsers). This approach allows modifications of the code that the author makes to be propagated instantly to any users simply by having the code updated at its home location. Users are guaranteed complete security for their computer environments. Finally, the speed of execution is, while not the fastest that can be achieved, certainly acceptable for this application. The ability to write this kind of research paper can and should change the way statistical research is performed.

REFERENCES

Bookstein, F. (1979). Fitting conic sections to scattered data. *Computer Graphics and Image Processing, 9*, 56–71.

Economist. (1995). Electronic Science Journals — Paperless Papers. *The Economist, Dec. 16*, 78–79.

Lehmann, E. (1983). *Theory of Point Estimation.* New York: John Wiley & Sons.

Rosin, P. (1993). A note on the least-squares fitting of ellipses. *Pattern Recognition Letters, 14*, 799–808.

Seber, G., & C. Nelder. (1989). *Nonlinear Regression.* New York: John Wiley.

Sellers, K., G. Chioran, S. Dain, . S. Benes, M. Lubos, K. Rammohan, & P. King-Smith. (1986). Red-green mixture thresholds in congenital and acquired color defects. *Vision Research, 26*(7), 1083–1097.

DISTRIBUTION LIST

1. Research Office (Code 09) 1
   Naval Postgraduate School
   Monterey, CA 93943-5000

2. Dudley Knox Library (Code 013) 2
   Naval Postgraduate School
   Monterey, CA 93943-5002

3. Defense Technical Information Center 2
   8725 John J. Kingman Rd., STE 0944
   Ft. Belvoir, VA 22060-6218

4. Therese Bilodeau 1
   Dept of Operations Research
   Naval Postgraduate School
   Monterey, CA 93943-5000

5. Prof. Samuel E. Buttrey (Code OR/Sb) 3
   Dept of Operations Research
   Naval Postgraduate School
   Monterey, CA 93943-5000

6. Prof. Gordon H. Bradley (Code OR/Bz) 2
   Naval Postgraduate School
   Monterey, CA 93943-5000

7. Air Force Office of Scientific Research 2
   ATTN: Dr. Neal Glassman (Code NM)
   110 Duncan Avenue, Suite B115
   Bolling AFB, DC 20332-8080

8. Office of Naval Research 2
   ATTN: Dr. Donald K. Wagner (Code 1111)
   800 North Quincy Street
   Arlington, VA 22217